DOCUMENT RESUME

ED 296 002 TM 011 922

AUTHOR Melican, Gerald J.; And Others
TITLE Accuracy of Option Elimination Prediction: Effect of a Technique To Improve Rater Agreement with Empirical Data.
PUB DATE 87
NOTE 21p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Certification; College Students; *Cutting Scores; *Difficulty Level; *Feedback; Higher Education; *Item Analysis; Licensing Examinations (Professions); Minimum Competency Testing; Multiple Choice Tests; Professors; Scoring; *Test Items
IDENTIFIERS *Nedelsky Method; Rater Reliability; Standard Setting

ABSTRACT

The effects of feedback about the ratings of other judges on subsequent ratings using the Nedelsky method and the ability of judges to retain or eliminate options in a manner consistent with the judgments of minimally competent examinees were studied using data from a basic algebra examination administered to 227 college students in 1987. The instrument used included 35 four-option multiple-choice items and was administered to the subjects at the start of an introductory statistics course to identify those in need of remedial instruction. Seven experienced instructors (judges) provided item ratings. Feedback did have an effect. Correlations between judges' estimates of item difficulty and the actual item difficulty increased slightly on the second rating. Improvements in both correlations and the accuracy of option elimination suggest that iterative procedures may be of value with the Nedelsky method. The requirement that judges evaluate individual options in a Nedelsky procedure can lead to accurate assessments of the plausibility of options. Nine tables present the study results. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED296002

Accuracy of Option Elimination Prediction: Effect of a Technique to Improve Rater Agreement with Empirical Data

Gerald J. Melican

Craig N. Mills

Educational Testing Service

Barbara S. Plake

and

Marie T. Benkofske

University of Nebraska

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
BARBARA S. PLAKE
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

In most certification/licensing examinations programs, cut-off scores are established using methods which rely upon expert judgment of item content such as the Angoff (1971) or Nedelsky (1954) methods. Although the methods have been widely used, the results are sometimes questioned for a variety of reasons, including low correlations between estimated and observed item difficulty and the tendency for judges to overestimate the ability level of examinees (or underestimate item difficulty).

One possible explanation for the discrepancies between judges' ratings and empirical data is that they have only one information source available during the rating task - the item content. Allowing judges to compare their ratings on every item and discuss discrepancies is one technique by which individuals can identify flaws in their judgments. However, such a practice can be time consuming and, if the judge group is large, cumbersome. In addition, there is the possibility that one dominant individual may control

TM011922

the discussion and skew the ratings. Provision of the ratings of other judges in an iterative process without discussion may represent an efficient procedure which does not suffer from the drawbacks of a group discussion.

Melican and Mills (1987) utilized this type of feedback with both the Angoff and Nedelsky methods for two certification tests. Judges rated each item twice.. On the second rating, information was provided about the ratings of the other judges. Correlations between estimated and observed item difficulty increased for the Angoff method. Increases were found for one test using the Nedelsky method, but correlations decreased for the second test. Melican and Mills cautioned, however, that small sample sizes limited the interpretability of their results. One purpose of this study was to investigate the effect of knowledge of other judges' ratings on the accuracy of subsequent ratings of the same items using the Nedelsky method.

The Nedelsky method requires judges to make a dichotomous judgement about each distractor for a multiple choice item. The judgement is whether or not the minimally competent examinee will be able to recognize the option as incorrect. An assumption is then made that the examinee will guess randomly among the remaining options. Thus, an estimated item difficulty for the item is calculated by taking the reciprocal of the number of options remaining.

For example, then estimated difficulty of a four-option multiple choice item for which one option is judged to be recognizably incorrect is 0.33.

In a recent study Melican, Mills, and Plake, (1987) tested the assumption that minimally competent examinees guess randomly among options judges perceive to be attractive. In that study, "functional" options were identified for each item. A functional option was one that was not identified as recognizably incorrect by the majority of the judges. If judges correctly identified distractors that were attractive to minimally competent candidates (MCCs) and MCCs guessed randomly among those distractors, the responses of MCCs should have been randomly distributed among the correct answer and the functional distractors. This result was not obtained. Further, the accuracy of predicted item performance appeared to vary as a function of item difficulty and content domain. Correlations between estimated item difficulty and observed item difficulty in the MCC group ranged from -.12 to .54 across content groupings. However, the data indicated that judges could provide some useful data about the attractiveness of options. A second purpose of this study was to investigate the degree to which judges eliminated options that minimally competent examinees also eliminated as clearly incorrect.

Method

Instrument. The instrument used in this study is designed to measure basic algebra skills. It consists of 35 four-option multiple choice items. The test is administered to students at the beginning of an introductory statistics course to identify those who are in need of remedial instruction. The KR20 reliability estimate of the instrument is .87.

Collection of Data from Judges. Seven individuals who had recent experience as instructors of the course provided the item ratings. The judges were sent an invitation letter detailing the focus of the study and asking them to participate. All judges agreed to participate in the study. The Nedelsky method was used to rate the items. Item reviews were completed independently. Several weeks later, the test was distributed to judges for a reevaluation. The copy of the test provided to each judge for the second review included (1) identification of the options which that judge had eliminated and (2) the number of judges eliminating each option during the first rating.

Collection of Data from Examinees. The examination was administered via computer to 227 students during the summer of 1987. All students were required to complete the examination as part of the course requirements. Students were informed that they would be required to receive remedial

instruction if they scored at or below the cut-off score. The cut-off score was not announced until all subjects had completed the test.

During the test administration, examinees were instructed to identify options they felt were incorrect before answering each item. These ratings were not collected on the computer, but were recorded on a separate form. Following the test, examinees whose scores fell within two conditional standard errors of measurement of the cut-off score were identified and classified as Empirically Minimally Competent Candidates (EMCCs). Ninety-four examinees fell into this category.

For each item, each option was placed into one of three mutually exclusive categories based upon the ratings supplied by the EMCCs. If one-third or less of the EMCCs eliminated the option, it was placed in Category 1 (plausible options). If more than one-third, but less than two-thirds of the EMCCs eliminated an option, it was placed in Category 2. Options which were eliminated by two-thirds or more of the EMCCs were placed in Category 3 (implausible options). Correct answers were included in all option analyses since EMCCs could eliminate them as plausible answers.

Analyses. Cut-off scores were computed for each of the judges and the total group. The interclass correlation was computed. To evaluate the effect of knowledge of other judges' ratings on the accuracy of subsequent ratings,

correlations between empirical item difficulty and the estimated item difficulty for each judge for both rating tasks were compared.

The degree to which judges were able to identify options as clearly incorrect was assessed by comparing their ratings to those provided by the EMCCs. A judge who accurately perceived the attractiveness of options to the EMCCs would retain options in Category 1, eliminate options in Category 3, and provide mixed judgments about those in Category 2. These comparisons, when viewed over the two rating occasions, also provide information about the effect of knowledge of other judges' ratings on accuracy of the second ratings.

Results

Accuracy of Item Difficulty Estimates

The results of the Nedelsky rating task on both iterations are contained in Table 1. Included are the cut-off score estimates of each judge for each rating and the difference between those estimates. Correlations between item difficulty estimates and the item difficulty in the EMCC group are reported in the bottom portion of the table as are the correlations between the estimates on the two ratings.

The absolute magnitude of the changes in cut-off scores for the individual judges ranged from .08 to 2.73 raw score points. For five of the seven judges, the cut-off score obtained for the second rating was adjusted in the direction of the mean cut-off of the first rating. That is, three of the four judges whose initial rating was below the average provided higher cut-offs on the second rating. Two of the three judges who were above the average lowered their cut-off scores on the second rating. The revisions to the cut-off scores ranged from a decrease of 1.10 to an increase of 2.73 raw score points. The average absolute change in the cut-off score was 1.01 raw score points (2.89 percent on the 35-item test). The average cut-off score increased by .73 points and the standard deviation of the ratings decreased by .54 points. The interclass correlation coefficient was .69 on the first occasion and .85 on the second.

Insert Table 1 About Here

More important than the reduction in variance or increase in agreement, however, is the relationship of the judges' ratings to the actual item difficulty. Correlations improved for five of the seven judges. The two judges for whom the correlation did not improve were the two whose cut off score adjustments were in the direction away from the mean of the first

rating. For the total group, the correlation improved from .67 to .71 on the second rating.

Accuracy of Option Ratings

To assess judges' abilities to recognize options that were attractive to EMCCs, their judgments were compared to the examinees' judgments using the categorization scheme described above. Judges' ratings were considered accurate if they retained options which were considered plausible by the EMCCs (Category 1) or eliminated options which were considered implausible by the EMCCs (Category 3). Reported in Tables 2a through 2g are the number of options eliminated and retained by the judge on the first and second ratings by EMCC category. The accuracy of ratings for each occasion is shown in the marginals. Accurate classifications are those in Category 1 which were retained by the judge and those in Category 3 which were eliminated. These accurate judgements are shown in **boldface** in the marginals. Inaccurate judgements are underlined in the marginals. The options reported in the upper left and lower right are those for which the judge provided consistent ratings on both occasions. Options in the upper right were eliminated on the first rating, but retained on the second. Improved ratings are those options in this cell classified as Category 1. Options in the lower left cell were retained on the first rating eliminated on the second. Category 3 options in this cell are those for which the rating improved.

Assessing accuracy in this manner eliminates options in Category 2 from consideration. This has the advantage of limiting the evaluation of a judge's accuracy to those options for which their was a clear majority in the EMCC group. Category 2 contained 48 options. Thus, each judge could make up to 92 correct classifications. On average, for these 92 options, judges made 68.9 correct decisions on the first rating and 72.9 on the second. An assessment of accuracy of the ratings of individual options in Category 2 cannot be made directly when judges make dichotomous decisions. If judges, however, are able to recognize options that will appear plausible to about half of the EMCCs, a reasonable assumption might be that they would retain about half of these options and eliminate the rest. In general, this result was not found and improvements for options in this category were not noted on the second rating. The tendency of the judges was to assign these options to the retain category.

Insert Tables 2a - 2g About Here

The accuracy of judgments within the total group of judges was assessed by comparing the number of judges eliminating each option with the EMCC category. These results are reported in Table 3. The axes in Table 3 are the number of judges eliminating options. Within each cell, options are

reported by EMCC category. These results reinforce the results for individual judges.

In general, most judges retained Category 1 options. Of the 33 options placed in Category 1, 23 were retained by all judges at Time 1 and 28 were retained at Time 2. Five were eliminated by one judge on the first ratings and by no judges on the second. Only two options were not retained by a clear majority of the judges on both ratings; one option was eliminated by three judges in both ratings and one was eliminated by four judges in both ratings.

Most judges eliminated Category 3 options, but there were more misclassifications of these options than of Category 1 options. Of the 59 options in this category, only 9 were eliminated by all judges on both ratings. Five or more judges eliminated an additional 25 options on both ratings. Ten of the remaining options were eliminated by four or more judges in both ratings. Fourteen Category 3 options were eliminated by three or fewer judges on both occasions.

Insert Table 3 About Here

Category 2 options were retained more often than they were eliminated. Of the 48 options in this category, 31 were eliminated by two or fewer judges on both occasions.

Summary and Discussion

The purposes of this study were to investigate (1) the effect of feedback on the ratings of other judges on subsequent ratings using the Nedelsky method and (2) the ability of judges to retain or eliminate options in a manner consistent with the judgments of minimally competent examinees. Feedback did have an effect and correlations between judges' estimates of item difficulty and actual item difficulty increased slightly on the second rating. However, the initial correlations between estimated and observed item difficulties were higher than those previously reported for the Nedelsky method (Melican and Mills, 1987, and Cross, Impara, Frary, and Jaeger, 1984) and the magnitude of the increase was not large. Nonetheless, the improvements in both correlations and the accuracy of option elimination for most judges suggest that iterative procedures may be of value with the Nedelsky method.

Judges were able to identify options that examinees retained. These results are expected since most judges and examinees retained the keyed response for most items. However, only 33 options were placed in Category 1 while there were 35 items on the test. Thus, at least two correct answers

were eliminated from consideration by more then one-third of the EMCCs. In this study, judges were not provided with correct answers and several judges eliminated the key for one or more items. These results suggest that, contrary to usual practice, individuals should consider obtaining item judgements without providing the correct answers to the judges, at least for the Nedelsky method.

In general, judges eliminated options that were eliminated by the EMCCs, but there was variance among the judgements about these options. Retaining these options, which EMCCs recognize as clearly incorrect has the effect of lowering the estimated item difficulty for the item and thus, produces a lower cut-off score than would be provided by the EMCCs themselves.

There was a tendency to retain options for which the EMCCs provided no clear judgement (Category 2 options). The effect of these judgments on Nedelsky ratings is to increase the number of options retained thereby decreasing the cut-off score. Retaining these options is, perhaps, to be expected. If judges perceive the options as potentially attractive to a significant portion of the EMCCs, it is reasonable to assume they would take a lenient approach (retain the option). A possible weakness of the Nedelsky method as it is typically applied may be the inability of judges to differentiate between clearly plausible, clearly implausible, and <u>possibly plausible</u> options. This issue could be addressed by developing different methods for calculating estimated cut-off scores using the data typically

collected for the Nedelsky method. Alternately, a modification of the Nedelsky method which allows judges to provide three ratings of option plausibility might be considered. Such a method, called the Minimally Acceptable Performance Level (MAPL) method, has been reported by Guerin and Smilansky (1976). In this method, options which are recognizably incorrect are given a value of 0, the correct answer and options which should not be recognized as incorrect as given a value of two, and the remaining options receive a value of 1. The estimated item difficulty is the probability that the minimally competent examinee would select the correct option by chance given the ratings. The results of this study suggest that further investigation of the MAPL is warranted.

There are several details of the procedures which limit the generalizability of this study. As noted previously, the correlations between estimated and empirical item difficulty were high. The process of collecting option elimination information from examinees while they were testing could have altered their test taking strategy so it conformed more closely to the judges' rating task. Further, the judges all had prior experience with students similar to those tested and may have had a more accurate perception of minimal competence than is usual in standard setting situations. Finally, rating mathematical items using the Nedelsky technique may be less difficult than rating other types of tests.

The results of this study, indicate that the requirement of the

Nedelsky procedure that judges evaluate individual options can lead to accurate assessments of the plausibility of options. Refinements of the manner in which Nedelsky ratings are collected and/or estimated item difficulty and cut-off scores are calculated may be warranted. Direct incorporation of the evaluation of options into the judgments of item difficulty with other judgmental methods, such as the Angoff method may also be of value.

Table 1
Cut-off Scores and Correlations
First and Second Ratings

	Judge 1	2	3	4	5	6	7	Total Mean	SD
Cut-off Scores									
Time 1	13.23	17.23	21.32	17.54	11.36	24.23	23.07	18.28	4.88
Time 2	15.96	18.39	20.89	18.38	11.28	23.13	23.81	18.83	4.34
T2 - T1	2.73	1.16	-.43	.84	-.08	-1.10	.74		
Correlations									
Time1, Observed	.66	.46	.43	.09	.34	.47	.64	.67	
Time2, Observed	.71	.61	.55	.51	.31	.60	.56	.71	
Time1, Time2	.64	.74	.80	.57	.86	.85	.63		

Table 2a
Classification of Options by EMCC Category
Within Judge Rating for Judge 1

	Rating Category	Time 1 Retain	Eliminate	Total Time 2
Time 2				
	1	33	0	33
Retain	2	41	0	41
	3	17	2	19
	1	0	0	0
Eliminate	2	6	1	7
	3	8	32	41
Total	1	33	0	33
Time 1	2	47	1	48
	3	25	34	59

Table 2b
Classification of Options by EMCC Category
Within Judge Rating for Judge 2

	Rating Category	Time 1 Retain	Eliminate	Total Time 2
Time 2				
Retain	1	31	1	32
	2	34	4	38
	3	19	1	20
Eliminate	1	0	1	1
	2	3	7	10
	3	7	32	39
Total Time 1	1	31	2	33
	2	37	11	48
	3	26	33	59

Table 2c
Classification of Options by EMCC Category
Within Judge Rating for Judge 3

	Rating Category	Time 1 Retain	Eliminate	Total Time 2
Time 2				
Retain	1	31	0	31
	2	35	2	37
	3	9	3	12
Eliminate	1	0	2	2
	2	2	9	11
	3	4	43	47
Total Time 1	1	31	2	33
	2	37	11	48
	3	13	46	59

Table 2d
Classification of Options by EMCC Category
Within Judge Rating for Judge 4

		Time 1 Rating Category	Retain	Eliminate	Total Time 2
Time 2	Retain	1	32	0	32
		2	28	3	31
		3	17	0	17
	Eliminate	1	0	1	1
		2	0	17	17
		3	6	36	42
	Total Time 1	1	32	1	33
		2	28	20	48
		3	23	36	59

Table 2e
Classification of Options by EMCC Category
Within Judge Rating for Judge 5

		Time 1 Rating Category	Retain	Eliminate	Total Time 2
Time 2	Retain	1	32	1	33
		2	41	2	43
		3	39	2	41
	Eliminate	1	0	0	0
		2	2	3	5
		3	2	16	18
	Total Time 1	1	32	1	33
		2	43	5	48
		3	41	18	59

Table 2f
Classification of Options by EMCC Category
Within Judge Rating for Judge 6

	Rating Category	Time 1 Retain	Eliminate	Total Time 2
Time 2	1	26	4	30
Retain	2	22	5	27
	3	9	0	9
	1	0	3	3
Eliminate	2	4	17	21
	3	1	49	50
Total	1	26	7	33
Time 1	2	26	22	48
	3	10	49	59

Table 2g
Classification of Options by EMCC Category
Within Judge Rating for Judge 7

	Rating Category	Time 1 Retain	Eliminate	Total Time 2
Time 2	1	30	0	30
Retain	2	19	9	28
	3	7	0	7
	1	1	2	3
Eliminate	2	4	16	20
	3	2	50	52
Total	1	31	2	33
Time 1	2	23	25	48
	3	9	50	59

Table 3
Classification of Options by EMCC Category
Within Number of Judges Eliminating Options
Time 1 and Time 2

Number of Judges Eliminating Option		TIME 1 EMCC Category	0	1	2	3	4	5	6	7	Total Time 2
		1	23	5							28
	0	2	9	10							19
		3	2	3							5
		1		3	0						3
	1	2		2	6						8
		3		0	2						2
		1		0	0	0					0
	2	2		1	3	2					6
		3		0	2	0					2
T											
I		1		0	0	1					1
M	3	2		0	0	4					4
E		3		1	0	4					5
2		1				0	1				1
	4	2				3	1				4
		3				0	6				6
		1				0	0	0	0		0
	5	2				0	1	1	0		2
		3				1	2	2	1		6
		1					0	0	0		0
	6	2					1	1	0		2
		3					1	11	6		18
		1					0	0	0	0	0
	7	2					0	3	0	0	3
		3					1	1	4	9	15
	Total	1	23	8	0	1	1	0	0	0	33
	Time 1	2	9	13	9	9	3	5	0	0	48
		3	2	4	4	5	10	14	11	9	59

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.) Educational Measurement, Washington, D.C.: American Council on Education.

Cross, L. H., Impara, J. C., Frary, R. B. & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21, 113-129.

Guerin, R. O. & Smilansky, J. (1976). The accuracy of absolute minimal acceptable performance levels for multiple-choice examinations. Journal of Medical Education, 51, 416-417.

Lorge, I., & Kruglov, L. K. (1953). The improvement of the estimates of tests difficulty. Educational and Psychological Measurement, 13, 34-46.

Melican, G. J. & Mills, C. N. (1987,April). The effect of knowledge of other judges' ratings of item difficulty in an iterative process using the Angoff and Nedelsky methods. A paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Melican, G. J., Mills, C. N., & Plake, B. S. (1987, April). Accuracy of item performance predictions based upon the Nedelsky standard setting method. A paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.